PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION CONCERNING SUBMISSION OR TRANSMITTAL OF PRIORITY DOCUMENT

(PCT Administrative Instructions, Section 411)

To:

KOIKE, Akira
No.11 Mori Bldg., 6-4, Toranomon 2-chome
Minato-ku, Tokyo 105-0001
JAPON

Date of mailing (day/month/year) 02 February 2000 (02.02.00)	
Applicant's or agent's file reference SK00PCT7	**IMPORTANT NOTIFICATION**
International application No. PCT/JP00/00203	International filing date (day/month/year) 18 January 2000 (18.01.00)
International publication date (day/month/year) Not yet published	Priority date (day/month/year) 21 January 1999 (21.01.99)
Applicant SONY CORPORATION et al	

1. The applicant is hereby notified of the date of receipt (except where the letters "NR" appear in the right-hand column) by the International Bureau of the priority document(s) relating to the earlier application(s) indicated below. Unless otherwise indicated by an asterisk appearing next to a date of receipt, or by the letters "NR", in the right-hand column, the priority document concerned was submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b).

2. This updates and replaces any previously issued notification concerning submission or transmittal of priority documents.

3. An **asterisk(*)** appearing next to a date of receipt, in the right-hand column, denotes a priority document submitted or transmitted to the International Bureau but not in compliance with Rule 17.1(a) or (b). In such a case, **the attention of the applicant is directed** to Rule 17.1(c) which provides that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity, upon entry into the national phase, to furnish the priority document within a time limit which is reasonable under the circumstances.

4. The **letters "NR"** appearing in the right-hand column denote a priority document which was not received by the International Bureau or which the applicant did not request the receiving Office to prepare and transmit to the International Bureau, as provided by Rule 17.1(a) or (b), respectively. In such a case, **the attention of the applicant is directed** to Rule 17.1(c) which provides that no designated Office may disregard the priority claim concerned before giving the applicant an opportunity, upon entry into the national phase, to furnish the priority document within a time limit which is reasonable under the circumstances.

Priority date	Priority application No.	Country or regional Office or PCT receiving Office	Date of receipt of priority document
21 Janu 1999 (21.01.99)	11/13307	JP	28 Janu 2000 (28.01.00)

The International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland

Facsimile No. (41-22) 740.14.35

Authorized officer

Y. KUWAHARA

Telephone No. (41-22) 338.83.38

PCT/JP00/00203

JP00/203

日本国特許庁
PATENT OFFICE
JAPANESE GOVERNMENT

18.01.00

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９９年　１月２１日

出願番号
Application Number: 平成１１年特許願第０１３３０７号

出願人
Applicant (s): ソニー株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

１９９９年１２月　３日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

出証番号　出証特平１１－３０８４５９８

【書類名】　特許願

【整理番号】　9801108806

【あて先】　特許庁長官殿

【国際特許分類】　G06F　7/00

【発明者】

【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内

【氏名】　長尾　確

【特許出願人】

【識別番号】　000002185

【氏名又は名称】　ソニー株式会社

【代表者】　出井　伸之

【代理人】

【識別番号】　100067736

【弁理士】

【氏名又は名称】　小池　晃

【選任した代理人】

【識別番号】　100086335

【弁理士】

【氏名又は名称】　田村　榮一

【選任した代理人】

【識別番号】　100096677

【弁理士】

【氏名又は名称】　伊賀　誠司

【手数料の表示】

【予納台帳番号】　019530

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9707387

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書処理方法および装置ならびに記録媒体

【特許請求の範囲】

【請求項１】　複数の要素から構成される内部構造を有する文書を処理する文書処理方法において、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出工程と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類工程と

を有することを特徴とする文書処理方法。

【請求項２】　複数の文書を受信する受信工程を有し、

上記特徴情報抽出工程は、上記受信工程で受信した各文書についてその文書の特徴を表す特徴情報を抽出すること

を特徴とする請求項１記載の文書処理方法。

【請求項３】　上記特徴情報抽出工程は、上記文書の内部構造に基づいて各要素に重みを付与し、この重みが所定値より大きい要素を抽出すること

を特徴とする請求項１記載の文書処理方法。

【請求項４】　上記文書分類工程は、複数の文書を分類項目に分類した分類操作に基づいて作成された分類モデルを用いて文書を分類すること

を特徴とする請求項１記載の文書処理方法。

【請求項５】　上記文書分類工程は、上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目の特徴情報との関連度を計算し、上記関連度が閾値を超えると、上記関連度が最大となる分類項目に文書を分類すること

を特徴とする請求項１記載の文書処理方法。

【請求項６】　上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目の特徴情報の関連度は、上記文書の特徴情報に含まれる固有名詞と上記分類項目の特徴情報に含まれる固有名詞とに共通する固有名詞の数と、上記文書の特徴情報に含まれる固有名詞以外の語義の上記分類項目に含まれる固有名詞以外の語

出証特平１１－３０８４５９８

義に対する語義の関連度の総和との線形結合であること

を特徴とする請求項５記載の文書処理方法。

【請求項７】　上記語義の関連度は、語義の間の参照関係の構造に基づいて各語義に重みを付与し、一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度とすることにより得られたものであること

を特徴とする請求項６記載の文書処理方法。

【請求項８】　上記語義の間の参照関係は、各語義について他の語義を参照する辞書を用いて作成されたこと

を特徴とする請求項７記載の文書処理方法。

【請求項９】　複数の要素から構成される内部構造を有する文書を処理する文書処理方法において、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与工程と、

上記重み付与工程で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算工程と

を有することを特徴とする文書処理方法。

【請求項１０】　上記重み付与工程は、各語義について他の語義を参照する辞書を用いて語義の参照関係を組織した参照関係の構造に基づいて重みを付与すること

を特徴とする請求項９記載の文書処理方法。

【請求項１１】　上記辞書の各語義にはその属性を示す属性情報が付与され、上記参照関係の構造は上記属性情報に基づいて組織されること

を特徴とする請求項１０記載の文書処理方法。

【請求項１２】　複数の要素から構成される内部構造を有する文書を処理する文書処理装置において、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出手段と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出手段で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類手段と

を有することを特徴とする文書処理装置。

【請求項１３】　複数の要素から構成される内部構造を有する文書を処理する文書処理装置において、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与手段と、

上記重み付与手段で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算手段と

を有することを特徴とする文書処理装置。

【請求項１４】　複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、上記文書処理プログラムは、

上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出処理と、

分類モデルを構成する複数の分類項目について、上記特徴情報抽出処理で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類処理と

を有することを特徴とする記録媒体。

【請求項１５】　複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、上記文書処理プログラムは、

語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与処理と、

上記重み付与処理で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算処理と

を有することを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、要素について内部構造を付与された文書を処理する文書処理方法および装置ならびに上記文書を処理するプログラムを記録された記録媒体に関し、詳しくは、文書に含まれる語義の関連度に基づいて文書を分類するような文書処理方法および装置、ならびに文書に含まれる語義の関連度に基づいて文書を分類するような文書処理のプログラムが記録されてなる記録媒体に関する。

【０００２】

【従来の技術】

従来、インターネットにおいて、ウィンドウ形式でハイパーテキスト型情報を提供するアプリケーションサービスとしてWWW（World Wide Web）が提供されている。

【０００３】

WWWは、文書の作成、公開または共有化の文書処理を実行し、新しいスタイルの文書の在り方を示したシステムである。しかし、文書の実際上の利用の観点からは、文書の内容に基づいた文書の分類や要約といった、WWWを越える高度な文書処理が求められている。このような高度な文書処理には、文書の内容の機械的な処理が不可欠である。

【０００４】

しかしながら、文書の内容の機械的な処理は、以下のような理由から依然として困難である。第１に、ハイパーテキストを記述する言語であるＨＴＭＬ（Hyper Text Markup Language）は、文書の表現については規定するが、文書の内容についてはほとんど規定しない。第２に、文書間に構成されたハイパーテキストの

ネットワークは、文書の読者にとって文書の内容を理解するために必ずしも利用しやすいものではない。第３に、一般に文章の著作者は読者の便宜を念頭に置かずに著作するが、文書の読者の便宜が著作者の便宜と調整されることはない。

【０００５】

このように、WWWは新しい文書の在り方を示した革新的なシステムであるが、文書を機械的に処理しないために、高度な文書処理を行うことができなかった。換言すると、高度な文書処理を実行するためには、文書を機械的に処理することが必要となる。

【０００６】

そこで、文書の機械的な処理を目標として、文書の機械的な処理を支援するシステムが自然言語研究の成果に基づいて開発されている。自然言語研究による文書処理の最初のステップとして、文書の著作者等による文書の内部構造についての属性情報、いわゆるタグの付与を前提とした、文書に付与されたタグを利用する機械的な文書処理が提案されている。

【０００７】

【発明が解決しようとする課題】

ところで、近年のコンピュータの普及や、ネットワーク化の進展に伴い、文章処理や、文書の内容に依存した索引などで、テキスト文書の作成、ラベル付け、変更などを行う文書処理の高機能化が求められている。このような高機能な文書処理を実現するためには、文書内における各語義の関連度に基づいた文書処理が必要となる。

【０００８】

本発明は、上述の実情に鑑みて提案されるものであって、文書内における語義の関連度に基づいた文書処理を行うような文書処理方法および装置、ならびに文書内における関連度に基づいた文書処理のプログラムが記録されてなる記録媒体を提供することを目的とする。

【０００９】

【課題を解決するための手段】

上述の課題を解決するために、本発明に係る文書処理方法は、複数の要素から

構成される内部構造を有する文書を処理する文書処理方法において、上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出工程と、分類モデルを構成する複数の分類項目について、上記特徴情報抽出工程で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類工程とを有するものである。

【００１０】

また、本発明に係る文書処理方法は、複数の要素から構成される内部構造を有する文書を処理する文書処理方法において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与工程と、上記重み付与工程で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算工程とを有するものである。

【００１１】

本発明に係る文書処理装置は、複数の要素から構成される内部構造を有する文書を処理する文書処理装置において、上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出手段と、分類モデルを構成する複数の分類項目について、上記特徴情報抽出手段で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類手段とを有するものである。

【００１２】

また、本発明に係る文書処理装置は、複数の要素から構成される内部構造を有する文書を処理する文書処理装置において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与手段と、上記重み付与手段で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算手段とを有するものである。

【００１３】

本発明に係る記録媒体は、複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、上記プログラムは、上記文書についてその文書の特徴を表す特徴情報を抽出する特徴情報抽出処理と、分類モデルを構成する複数の分類項目について、上記特徴情報抽出処理で抽出した文書の特徴情報と上記分類項目毎の特徴情報との関連度に応じて、各文書を上記分類項目に分類する文書分類処理とを有するものである。

【００１４】

また、本発明に係る記録媒体は、複数の要素から構成される内部構造を有する文書を処理する文書処理プログラムが記録された記録媒体において、語義の間の参照関係の構造に基づいて各要素に重みを付与する重み付与処理と、上記重み付与処理で一の語義に付与された重みを独立変数とするとともに他の語義に付与された重みを従属変数とし、上記他の語義に付与された重みの差分を上記一の語義に付与された重みの差分で除した商を上記一の語義と上記他の語義の関連度として演算する関連度演算処理とを有するものである。

【００１５】

【発明の実施の形態】

以下、図面を参照して、本発明に係る文書処理方法および装置ならびに記録媒体の実施の形態について説明する。

【００１６】

本発明の実施の形態としての文書処理装置は、図１に示すように、制御部１１およびインターフェース１２を備える本体１０と、ユーザからの入力を受け付けて本体１０に送る入力部２０と、外部からの信号を受信して本体１０に送る受信部２１と、本体１０からの出力を表示する表示部３０と、記録媒体３２に対して情報を記録／再生する記録／再生部３１２とを有している。

【００１７】

本体１０は、制御部１１およびインターフェース１２を有し、この文書処理装置の主要な部分を構成している。制御部１１は、この文書処理装置における処理を集中して実行するＣＰＵ１３と、揮発性のメモリであるＲＡＭ１４と、不揮発

性のメモリであるＲＯＭ１５とを有している。ＣＰＵ１３は、たとえばＲＯＭ１４に記録された手順にしたがって、必要な場合にはデータを一時的にＲＡＭ１４に格納して、プログラムを実行するための制御を行う。インターフェース１２には、入力部２０、受信部２１および表示部３０が接続される。インターフェース１２は、制御部１１からの制御の下に、入力部２０および受信部２１からのデータの入力、表示部３０へのデータの送信について、データを送信するタイミングを調整したり、データの形式を変換したりする。

【００１８】

入力部２０は、この文書処理装置に対するユーザの入力を受け付ける部分である。この入力部２０は、たとえばキーボードやマウスにより構成される。ユーザは、この入力部２０を用い、キーボードによりキーワードを入力したり、マウスにより表示部３０に表示されている文書のエレメントを選択して入力したりすることができる。ここで、エレメントとは文書を構成する要素であって、たとえば文書、文および語が含まれる。

【００１９】

受信部２１は、この文書処理装置に外部からたとえば通信回線を介して送信される信号を受信する部分である。この受信部２１は、たとえば電子文書である複数の文書を受信する。受信部２１は、受信したデータを本体１０に送る。

【００２０】

出力部３０は、この文書処理装置からの出力結果を表示するものである。この出力部３０は、たとえば陰極線管（cathode ray tube;CRT）や液晶表示装置（liquid crystal display;LCD）から構成され、たとえば単数または複数のウィンドウを表示し、このウィンドウ上に文字、図形等を表示したりする。

【００２１】

記録／再生部３１は、この文書処理装置の制御部１１の制御の下に、たとえばいわゆるフロッピーディスクのような記録媒体３２に対して情報の記録／再生を行う。記録媒体３２には、たとえば文書の語義に基づいて関連度を求め、この関連度に基づいて文書処理を実行するようなプログラムが記録されている。なお、この記録媒体３２についてはさらに後述する。

【００２２】

続いて、本実施の形態における文書について説明する。本実施の形態における文書は、ツリー状のタグ付けによる内部構造を有している。本実施の形態においては、図２に示すように、このタグ付けによる内部構造、文書、文、語彙エレメント等の各エレメント、通常リンク、参照・被参照リンク等が、タグとしてあらかじめ文書に付与されている。図中において、白丸“○”は文書の要素すなわちエレメントであり、最下位の白丸は文書における最小レベルの語に対応する、語彙エレメントである。また、実線は語、句、節、文等の文書の構造を示す通常リンク（normal link ）である。破線は参照・被参照による係り受け関係を示す参照リンク（reference link）である。文書のタグ付けによる内部構造は、上位から下位への順序で、文書（document）、文書の下位であり段落の上位であるオプションのサブディビジョン（subdivision ）、オプションの段落（paragraph）、文（sentence ）、文の下位であるサブセンテンシャルセグメント（subsentential segment ）、・・・、最下位の語彙エレメントのような階層構造である。

【００２３】

本実施の形態においては、文書のタグ付けによる内部構造として、多言語間に共通な意味的・語用論的タグを文書に付与することにより、文書の機械的な内容理解を可能にするようなタグ付けを採用している。タグとは、データに対してその属性を表すために付加される属性情報である。

【００２４】

本実施の形態において文書のタグ付けによる内部構造は、ＨＴＭＬ（Hyper Text Markup Language）と同様にＸＭＬ（Extended Markup Language）の形式のタグである。すなわち、タグは、係り受け、たとえば代名詞の指示対象、多義語の意味のように統語（syntactic）・意味（semantic）等の情報を含んでいる。

【００２５】

文章のタグ付けによる内部構造の一例を次に示すが、文章へのタグ付けはこの方法に限られないことはもちろんである。

【００２６】

たとえば、“Time flies like an arrow.”という文については、

＜文＞＜名詞句　語義＝“time0”＞time＜／名詞句＞
＜動詞句＞＜動詞　語義＝“fly1”＞flies＜／動詞＞
＜形容動詞句＞＜形容動詞　語義＝like0＞like＜／形容動詞＞ ＜名詞句＞an
＜名詞　語義＝“arrow0”＞arrow＜／名詞＞＜／名詞句＞
＜／形容動詞句＞＜／動詞句＞.＜／文＞
というようにタグ付けすることができる。ここで＜文＞、＜名詞＞、＜名詞句＞、＜動詞＞、＜動詞句＞、＜形容動詞＞、＜形容動詞句＞は、それぞれ一文、名詞、名詞句、動詞、動詞句、前置詞句、後置詞句を含む形容詞／形容詞句、形容詞句／形容動詞句のような文の統語構造（syntactic structure ）を表している。タグは、エレメントの先端の前および終端の後に対応して配置される。ここでは、エレメントの終端の後ろに配置されるタグは、記号“／”により、文書の最小単位の要素、すなわちエレメントの終端であることを示している。このエレメントは統語的構成素、すなわち句、節、および文を示す。なお、語義＝time0は、語timeの有する複数の意味、すなわち複数の語義のうちの第０番目の意味を指している。具体的には、timeには名詞と動詞があるが、ここではtimeが名詞であることを示している。たとえば、語“オレンジ”は色と果物の意味があるが、これらも語義によって区別することができる。

【００２７】

先に図２で説明したような文書のタグ付けによる内部構造は、図３のウインドウ１０１に示すように、その統語構造を表示することができる。このウインドウ１０１においては、右半分１０３が語彙エレメントを、左半分１０２が文の構造を示している。

【００２８】

このウィンドウ１０１には、タグ付けされた次に示すような文書が表示されている。この文書においても、タグによって統語構造が記述されている。次に示す文書は、「Ａ氏のＢ会が終わったＣ市で、一部の大衆紙と一般紙がその写真報道を自主規制する方針を紙面で明らかにした。」についてのタグ付けによる内部構造を示すものである。

【００２９】

＜文書＞＜文＞＜形容動詞句　関係＝“場所”＞＜名詞句＞＜形容動詞句　場所＝“Ｃ市”＞
＜形容動詞句　関係＝“主語”＞＜名詞句 識別子＝“Ｂ会”＞＜形容動詞句　関係　“位置”＞Ａ氏の＜／形容動詞句＞Ｂ会＜／名詞句＞が＜／形容動詞句＞終わった＜／形容動詞句＞＜地名　識別子＝“Ｃ市”＞Ｃ市＜／地名＞＜／名詞句＞で、＜／形容動詞句＞＜形容動詞句　関係＝“主語”＞＜名詞句　識別子＝新聞”統語＝“並列”＞＜名詞句＞＜形容動詞句＞一部の＜／形容動詞句＞大衆紙＜／名詞句＞と＜名詞＞一般紙＜／名詞＞＜／名詞句＞が＜／形容動詞句＞
＜形容動詞句　関係＝“目的語”＞＜形容動詞句　関係＝“内容”　主語＝“新聞”＞＜形容動詞句　関係＝“目的語”＞＜名詞句＞＜形容動詞句＞＜名詞　共参照＝“Ｂ”＞そ＜／名詞＞の＜／形容動詞句＞写真報道＜／名詞句＞を＜／形容動詞句＞
自主規制する＜／形容動詞句＞方針を＜／形容動詞句＞
＜形容動詞句　関係＝“場所”＞紙面で＜／形容動詞句＞
明らかにした。＜／文＞＜／文書＞

【００３０】

この文章においては、「一部の大衆紙と一般紙」のように、統語＝“並列”は並列を表す。並列の定義は、係り受け関係を共有すると言うことである。特に何も指定がない場合は、たとえば、＜名詞句　関係＝ｘ＞＜名詞＞Ａ＜／名詞＞＜名詞＞Ｂ＜／名詞＞＜／名詞句＞ はＡがＢに依存関係のあることを表す。また、関係＝ｘ はこの＜名詞句＞エレメントの関係属性を表している。

【００３１】

続いて、タグ付けにおける、統語、意味、修辞についての相互関係を記述する関係属性について説明する。

【００３２】

主語、目的語、間接目的語のような文法機能、動作主、被動作者、受益者などのような主題役割、および理由、結果などのような修辞関係はこの関係属性によって表示する。関係属性は関係＝*** という形で表される。本実施の形態では、

比較的容易な文法機能、すなわち、主語、目的語、間接目的語のような文における当該語の機能について関係属性を記述する。

【００３３】

続いて、この文書処理装置の動作について、図４のフローチャートを参照して説明する。文書処理装置は、複数の文書について各文書の内容に関する特徴を表す特徴情報を含み、その文書の指標となるインデックスを作成する。そして、文書分類の分類モデルに基づいて、それぞれの文書のインデックスを参照することにより文書の自動的な分類を行う。分類モデルは、文書を分類する複数の分類項目から構成され、各分類項目は特徴を表す特徴情報を有している。

【００３４】

最初のステップＳ１１においては、文書処理装置の受信部２１は、外部から送信される複数の文書を受信する。文書処理装置は、受信部２１にて受信された複数の文書を、制御部１１の制御の下に、たとえばＲＡＭ１４や記録／再生部３１に記録する。文書は、図２に示したように、複数の要素すなわちエレメントからツリー状に構造化されたタグ付けによる内部構造を有している。

【００３５】

ステップＳ１２においては、ユーザは、文書処理装置の表示部３０に表示される文書を閲覧する。すなわち、文書処理装置の制御部１１は、ユーザの希望に応じて、記憶する複数の文書から表示部３０に文書を表示するように表示部３０を制御する。文書処理装置の表示部３０に表示する文書は、ユーザが入力部２０に上記複数の文書のうちから所望の文書の選択を入力することにより任意に選択される。表示部３０には、ユーザにより選択された文書の一部または全部の内容が、たとえばその領域の大きさを変更可能なウィンドウにより表示される。なお、ユーザが文書閲覧を行うこのステップＳ１２は、ユーザの必要に応じて設けられる。また、図中においてこのステップＳ１２が平行四辺形で表されているのは、ユーザが操作することに対応している。ステップＳ１３においても同様である。

【００３６】

ここで、表示部３０における表示の具体例について説明する。この具体例とは、ユーザが自由に文書を分類する分類項目であるカテゴリを設定、変更できるよ

うにしたものである。この具体例においては、ユーザが設定したカテゴリに応じて文書の自動分類が行われる。

【００３７】

このようなグラフィックユーザインターフェース（graphic user interface; GUI）の具体例は、図５に示すようになる。このＧＵＩ画面においては、操作ボタン３０２、“他のトピックス”を表示する第１の分類表示部３０３、“ビジネスニュース”を表示する第２の分類表示部３０４、“政治ニュース”を表示する第３の分類表示部３０５などをの各分類項目が表示されている。“他のトピックス”は、“他のトピックス”、“ビジネスニュース”等の特定の分類項目に分類していない文書が分類される。各分類項目の表示部においては、文書のタイトルや文書の最初の部分が表示される。

【００３８】

また、このＧＵＩにおいては、操作ボタン３０２は、画面のウィンドウの状態を初期の位置に戻すポジションリセット（position reset）と、文書の内容を閲読するブラウザ（browser ）を呼び出すブラウザのボタンと、このウィンドウから脱出するエグジット（exit）のボタンとを含んでいる。なお、上述の各分類表示部の大きさは固定的ではなく、所望の大きさに変更することができる。また、分類表示部のタイトルも自由に変更することができる。

【００３９】

この自動分類は、ユーザの個別の要求に応じて分類項目を決めることにより、ユーザの関心に応えたり、ユーザが文書を探すときの効率の向上を図るものである。

【００４０】

ステップＳ１３においては、ユーザは、ステップＳ１２において文書処理装置の表示部３０にて閲覧した複数の文書について、この複数の文書を分類する分類項目、いわゆるカテゴリを作成し、この分類項目に上記複数の文書を分類する。文書処理装置においては、文書を分類する分類項目の設定は、たとえば分類項目の数に対応して分割された領域を有するウインドウついて、所望の分類項目を追加したり、あるいは変更や削除をすることにより行われる。複数の文書の分類項

目への分類は、たとえば文書の一部や分類項目のタイトルが表示され、このタイトルに対応する領域が設けられたウインドウにおいて、たとえば画面上に表示されたアイコンをカーソルに連動するマウスをクリックしてドラッグすることにより、文書を所望の領域に移動して各文書を分類する。このような、文書の分類作成および分類操作は、ユーザが表示部３０の表示を参照しながら入力部２０に入力することにより行われる。作成された分類項目および分類操作の結果は、制御部１１の制御の下に、たとえばＲＡＭ１４に記録される。なお、文書の分類項目の作成および文書の分類の操作の詳細については、さらに後述する。

【００４１】

ステップＳ１４においては、文書処理装置の制御部１１は、ステップＳ１３において行われた分類項目の作成と、この分類項目に応じた分類操作に基づいて、分類モデルの作成を実行する。文書処理装置は、たとえばＲＡＭ１４に記録されたステップＳ１３における分類項目および分類操作の結果を読み出す。そして、文書処理装置の制御部１１は、この結果に基づいて、各分類項目に分類された上記複数の文書について、各分類項目に特徴的な固有名詞、固有名詞以外の語義、分類された文書のアドレスを集めて、分類モデルを生成する。ここで、固有名詞以外の場合に語そのものではなく語義を用いるのは、同じ語でも複数の意味を有することがあるからである。そして、文書処理装置の制御部１１は、このように作成した分類モデルをたとえばＲＡＭ１４に記憶する。なお、分類モデルの作成の詳細については、さらに後述する。

【００４２】

以上の一連の行程により、文書を分類する基準となる分類モデルが作成された。文書処理装置は、この分類モデルを基準として、文書を自動的に分類することができる。文書処理装置が行う、新たに受信した文書の自動的な分類の動作について、図６を参照して説明する。

【００４３】

文書処置装置において、外部からたとえば通信回線を介して受信部２１に新たな文書が送信されると、文書処理装置はこの文書を受信する。文書処理装置における文書の受信の動作については、上述したステップＳ１１で詳しく述べたので

、ここでは説明を省略する。受信した文書は、たとえばＲＡＭ１４や記録／再生部３１に記録される。

【００４４】

ステップＳ２２においては、文書処理装置の制御部１１は、たとえばＲＡＭ１４や記録／再生部３１に記録されたステップＳ２１で受信した文書を読み出す。制御部１１は、この新たな文書から各文書の特徴を表す語を抽出することによりその文書の指標、すなわちインデックスの作成を行う。そして、制御部１１は、各文書についてのインデックスをたとえばＲＡＭ１４に記録する。なお、このインデックス作成の詳細については、さらに後述する。

【００４５】

ステップＳ２３においては、文書処理装置の制御部１１は、分類モデルに基づいて、インデックスを附された各文書を上述のステップＳ１３において作成した複数の分類項目の一つに分類する。そして、制御部１１は、分類の結果をたとえばＲＡＭ１４に記録する。なお、このような文書の自動分類の詳細については、さらに後述する。

【００４６】

ステップＳ２４においては、文書処理装置の制御部１１は、たとえばＲＡＭ１４に記録されたステップＳ２３での新たな文書の自動分類の結果に基づいて、分類モデルを更新する。制御部１１は、更新した分類モデルをたとえばＲＡＭ１４に記録する。

【００４７】

上述したタグ付けによる内部構造を有する文書は、文書処理装置の受信部２１に外部から送信される。この文書は、たとえばデジタル符号化されたいわゆる電子文書である。文書処理装置は、このような文書をたとえばＲＡＭ１４や記録／再生部３１に記録する。ユーザは、文書処理装置の記録する複数の文書から、任意の文書を表示部３０に表示して閲読することができる。

【００４８】

表示部３０における文書の表示は、たとえば大きさを変更することができるウィンドウ上に表示することができる。また、文書の表示と共に、または文書の表

示に代えて要約を表示することができる。さらに、複数の文書をウィンドウにより並べて表示したり、複数のウィンドウを重ねて表示することができる。

【００４９】

文書処理装置の制御部１１は、このように表示部３０に表示された文書について、ユーザの入力にしたがい、各種の処理を実行する。ユーザによる、文書についての入力は、表示部３０に表示されるカーソルに連動する入力部２０のマウスをクリックすることにより表示部３０における所定の領域を指定したり、入力部２０のキーボードによりキーワードを入力したりすることにより行う。

【００５０】

次に、図４および図６の処理の詳細について説明する。

【００５１】

最初に、Ｓ２２の文書の特徴を発見してインデックスを作る手順について詳細に説明する。このインデックスとは、文書の特徴を表す語を各文書について抽出して指標としたものである。文書の特徴を発見してインデックスを作成する手順は、文書処理装置の制御部１１の制御の下に、図７のフローチャートに示す一連の手順により実行される。すなわち、以下の手順は語義の関連度を算出して、この関連度に基づいてインデックスの作成を行うものである。

【００５２】

最初のステップＳ３１においては、制御部１１は、図６のステップＳ２１において受信した文書についてその文書内で活性拡散を実行し、文書内の各エレメントの活性値を拡散する。制御部１１による文書の各エレメントへの活性値の拡散処理は、後に詳細に説明する。制御部１１は、活性拡散の結果として得られた活性値をたとえばＲＡＭ１４に記録する。

【００５３】

ステップＳ３２においては、制御部１１は、ステップＳ１１で得られた活性値に基づいて、活性値が予め設定された閾値を超えるエレメントを抽出する。制御部１１は、このように抽出したエレメントをたとえばＲＡＭ１４に記録する。

【００５４】

ステップＳ３３においては、制御部１１は、たとえばＲＡＭ１４からステップ

Ｓ３２にて抽出したエレメントを読み出す。そして、制御部１１は、このエレメントからすべての固有名詞を取り出してインデックスに加える。固有名詞は語義を持たず、辞書に載っていないなどの特殊の性質を有するので、固有名詞以外の語とは別に扱う。固有名詞であるか否かは、たとえば文書に付加されたタグにより識別される。たとえば、図３に示したタグ付けによる内部構造においては、“Ａ氏”、“Ｂ会”および“Ｃ市”は固有名詞である。そして、制御部１１は、取り出した固有名詞をインデックスに加え、その結果をたとえばＲＡＭ１４に記録する。

【００５５】

ステップＳ３４においては、制御部１１は、たとえばＲＡＭ１４からステップＳ３２にて抽出したエレメントから固有名詞以外の語義を取り出してインデックスに加え、その結果をＲＡＭ１４に記録する。ここでの語義とは、語の有する複数の意味の内の選択された一つである。本実施の形態においては、語義についてもタグ付けにより記述されている。

【００５６】

このように、文書の特徴を発見してインデックスを作成する手順は、複数のエレメントから構成されるタグ付けによる内部構造を有する文書の特徴を発見して、その特徴を配列したインデックスを作るものである。すなわち、タグ付けによる内部構造に基づいて上記文書について活性拡散をすることにより、各語彙エレメントの活性値を拡散し、拡散後の活性値が所定の閾値より大きい語彙エレメントを抽出する。そして、その語彙エレメントについて、固有名詞または語義をインデックスに追加する。

【００５７】

なお、上述のインデックスには、文書の特徴を表す語と共に、その文書のアドレスを含めることもできる。すなわち、インデックスはその文書を代表するような特徴を表す語を含むので、所望の文書を参照する際の指標とすることができる。このインデックスは、上述したように文書の自動的な分類に利用することができるが、これについての詳細はさらに後述する。

【００５８】

ここで、インデックスの具体例を示す。

【００５９】

＜インデックス　日付＝“AAAA/BB/CC”　時刻＝“DD:EE:FF”　文書アドレス＝“1234”＞

＜要約＞減税規模、触れず－Ｘ首相の会見＜／要約＞

＜語　語義＝“0003”　活性値＝“140.6”＞触れず＜／語＞

＜語　語義＝“0105”　識別子＝“Ｘ”　活性値＝“140.6”＞首相＜／語＞

＜名前　識別子＝“Ｘ”　語　語義＝“6103”　活性値＝“140.6”＞Ｘ首相＜／語＞

＜語　語義＝“5301”　活性値＝“140.6”＞求めた＜／語＞

＜語　語義＝“2350”　識別子＝“Ｘ”　活性値＝“140.6”＞首相＜／語＞

＜語　語義＝“9582”　活性値＝“140.6”＞強調した＜／語＞

＜語　語義＝“2595”　活性値＝“140.6”＞触れる＜／語＞

＜語　語義＝“9472”　活性値＝“140.6”＞予告した＜／語＞

＜語　語義＝“4934”　活性値＝“140.6”＞触れなかった＜／語＞

＜語　語義＝“0178”　活性値＝“140.6”＞釈明した＜／語＞

＜語　語義＝“7248”　識別子＝“Ｘ”　活性値＝“140.6”＞私＜／語＞

＜語　語義＝“3684”　識別子＝“Ｘ”　活性値＝“140.6”＞首相＜／語＞

＜語　語義＝“1824”　活性値＝“140.6”＞訴えた＜／語＞

＜語　語義＝“7289”　活性値＝“140.6”＞見せた＜／語＞

＜／インデックス＞

【００６０】

このインデックスにおいては、＜インデックス＞および＜／インデックス＞は、インデックスの始端および終端を、＜日付＞および＜時刻＞はこのインデックスが作成された日付および時刻を、＜要約＞および＜／要約＞はこのインデックスの内容の要約の始端および終端を示している。また、＜語＞および＜／語＞は語の始端および終端を、それぞれ示している。語義＝“0003”は、その語義が、複数の語義のうちの第３番目であることを示している。他についても同様である。

【００６１】

続いて、タグ付けによる内部構造に基づいて、ステップＳ３１で行う活性拡散によりエレメントの活性値を拡散する方法について説明する。なお、この活性拡散は、後述する図１１におけるステップＳ６２においても実行される。

【００６２】

タグ付けによる内部構造を与えられた文書においては、活性拡散と呼ばれる処理を行うことにより、各エレメントにタグ付けによる内部構造に応じた活性値を付与することができる。活性拡散は、活性値の高いエレメントと関わりのあるエレメントにも高い活性値を与えるような処理である。この活性値は、タグ付けによる内部構造に応じて決定されるので、タグ付けによる内部構造を考慮した文書の分析に利用することができる。

【００６３】

活性拡散は、図８のフローチャートに示す一連の行程にしたがって、文書処理装置の制御部１１の制御の下に実行される。

【００６４】

最初のステップＳ４１においては、文書内のエレメントの活性値の初期化が行われれる。すなわち、制御部１１は、語彙エレメントを除いたすべてのエレメントと語彙エレメントに対して活性値の初期値を割り当てる。たとえば、活性値の初期値として語彙エレメントを除いたすべてのエレメントに１を、語彙エレメントに零を、それぞれ割り当てればよい。また、制御部１１は、各エレメントの活性値の初期値に均一ではない値を割り当てることにより、活性拡散の結果得られた活性値の初期値の偏りを反映することができる。たとえば、ユーザが関心を有するエレメントに対しては、活性値の初期値を高く設定することにより、ユーザの関心を反映した活性値を得ることができる。

【００６５】

参照・被参照関係のエレメントを連結する参照・被参照リンクとそれ以外の通常リンクに関しては、エレメントを連結するリンクの端点の活性値である端点活性値を０に設定する。制御部１１は、このように付与した活性値の初期値をたとえばＲＡＭ１４に記録する。

【００６６】

エレメントとエレメントの連結は、たとえば図９に示すようになる。この図においては、文書を構成するエレメントとリンクの構造の一部として、エレメントE_iおよびエレメントE_jが示されている。エレメントE_iとエレメントE_jとは、活性値e_iおよびe_jをそれぞれ有し、リンクL_{ij}にて接続されている。リンクL_{ij}のエレメントE_iに接続する端点はT_{ij}、エレメントE_jに接続する端点はT_{ji}である。エレメントE_iは、リンクL_{ij}により接続されるエレメントE_jの他に、リンクL_{ik}、L_{il}およびL_{im}によって図示しないエレメントE_k、E_lおよびE_mにそれぞれ接続している。エレメントE_jは、リンクL_{ji}により接続されるエレメントE_iの他に、リンクL_{jp}、L_{jq}およびL_{jr}によって図示しないエレメントE_p、E_qおよびE_rにそれぞれ接続している。

【００６７】

ステップＳ４２においては、文書処理装置の制御部１１は、文書を構成するエレメントE_iを計数するカウンタの初期化を行う。すなわち、エレメントを計数するカウンタのカウント値ｉを１に設定する。すなわち、このカウンタは、第１番目のエレメントE_1を参照している。

【００６８】

ステップＳ４３においては、文書処理装置の制御部１１は、カウンタが参照するエレメントについて、活性値を計算するリンク処理を実行する。このリンク処理については、さらに後述する。

【００６９】

ステップＳ４４においては、文書処理装置の制御部１１は、文書中のすべてのエレメントについて活性値の計算が完了したか否かを判断する。そして、制御部１１は、文書中のすべてのエレメントについて活性値の計算が完了したときには“ＹＥＳ”としてステップＳ４５に処理を進め、文書中のすべてのエレメントについて活性値の計算が完了していないときには“ＮＯ”としてステップＳ４７に処理を進める。

【００７０】

具体的には、制御部１１は、カウンタにて計数されている活性値の計算がなさ

れたエレメントを参照するカウンタのカウント値ｉが、文書の含むエレメントの総数に達したか否かを判断する。そして、制御部１１は、カウンタのカウント値ｉが文書に含まれるエレメントの総数に達したときには、すべてのエレメントが計算済みとしてステップＳ４５に処理を進め、カウンタのカウント値ｉが文書に含まれるエレメントの総数に達していないときにはすべてのエレメントについて計算が終了していないとしてステップＳ４７に処理を進める。

【００７１】

ステップＳ４７においては、文書処理装置の制御部１１は、カウンタのカウント値ｉを１増加させて、カウンタのカウント値をｉ＋１とする。このことにより、カウンタはｉ＋１番目のエレメント、すなわち次のエレメントを参照する。そして、処理はステップＳ４３に戻り、端点活性値の計算およびこれに続く一連の行程が、次のｉ＋１番目のエレメントについて実行される。

【００７２】

具体的には、制御部１１は、エレメントを計数するカウンタのカウント値ｉを１増加する。このことにより、カウンタはステップＳ４３で活性値が計算されたエレメントの次のエレメントを参照することになる。

【００７３】

ステップＳ４５においては、文書処理装置の制御部１１は、文書に含まれるすべてのエレメントの活性値の変化分、すなわち新たに計算された活性値の元の活性値に対する変化分について、文書に含まれるすべてのエレメントについて平均値を計算する。

【００７４】

文書処理装置の制御部１１は、たとえばＲＡＭ１４に記録された元の活性値と新たに計算した活性値を、文書に含まれるすべてのエレメントについて読み出す。制御部１１は、新たに計算した活性値の元の活性値に対するそれぞれの変化分の総和を文書に含まれるエレメントの総数で除することにより、すべてのエレメントの活性値の変化分の平均値を計算する。制御部１１は、このように計算したすべてのエレメントの活性値の変化分の平均値を、たとえばＲＡＭ１４に記録する。

【００７５】

ステップＳ４６においては、制御部１１は、ステップＳ４９で計算したすべてのエレメントの活性値の変化分の平均値が、予め設定された閾値以内であるか否かを判断する。そして、制御部１１は、上記変化分が閾値以内であると“ＹＥＳ”としてこの一連の行程を終了する。上記制御部１１は、上記変化分が閾値以内でないときには“ＮＯ”として、ステップＳ４２にてカウンタのカウント値ｉを１に設定して文書のエレメントの活性値を計算する一連の行程を再び実行する。この一連の行程にて構成されるステップＳ４２からステップＳ４４に至るループが繰り返される毎に上記変化分は徐々に減少する。

【００７６】

続いて、ステップＳ４３にて実行される活性値を計算するリンク処理について、図１０に示すフローチャートを参照して説明する。

【００７７】

ステップＳ５１においては、文書処理装置の制御部１１は、図９に示すように、文書を構成するエレメントE_jを計数するカウンタの初期化を行う。すなわち、エレメントを計数するカウンタのカウント値ｊを１に設定する。すなわち、このカウンタは、第１番目のエレメントE_jを参照している。

【００７８】

ステップＳ５２においては、エレメントE_iとE_jを接続するリンクL_{ij}においては、制御部１１は、タグを参照することにより、そのリンクL_{ij}が通常リンクであるか否かを判断する。制御部１１は、リンクL_{ij}について、そのリンクが、語に対応する語彙エレメント、文に対応する文エレメント、段落に対応する段落エレメントなどの間の関係を示す通常リンクと、参照・被参照による係り受けの関係を示す参照リンクのいずれであるかを判断する。これは、図３の“関係”を参照することで判断することができる。制御部１１は、そのリンクが通常リンクのときには“ＹＥＳ”としてステップＳ５３に処理を進め、そのリンクが参照リンクのときには“ＮＯ”としてステップＳ５４に処理を進める。

【００７９】

ステップＳ５３においては、通常リンクL_{ij}に対して、そのリンクの端点の活

性値を計算する処理が行われる。この端点活性値の計算について、図９を参照して説明する。

【００８０】

ここでは、ステップＳ５２における判別により、リンクL_{ij}は通常リンクであることが明らかになっている。通常リンクL_{ij}に関して、エレメントE_iに接続する端点T_{ij}の端点活性値t_{ij}は、このリンクL_{ij}を除いたエレメントE_iに接続するすべてのリンクL_{ik}、L_{il}およびL_{im}の端点活性値t_{ik}、t_{il}およびt_{im}と、このエレメントE_iがリンクL_{ij}により接続するエレメントE_jの活性値e_jを加算し、この加算で得た値を文書に含まれるエレメントの総数で除することにより求められる。

【００８１】

エレメントE_{ij}の端点T_{ij}の端点活性値は、端点T_{ij}を一端とするリンクL_{ij}が通常リンクの場合、リンクL_{ij}の他端が接続されているエレメントE_jの端点の端点活性値のうちそのリンクL_{ij}と接続されている端点T_{ji}を除いたすべての端点の端点活性値、およびそのリンクL_{ij}が接続されるエレメントE_jの活性値e_jの和を文書全体に含まれるエレメントの総数で除することにより得られる。このような手順により、活性拡散における活性値の収束が保証されることになる。

【００８２】

文書処理装置の制御部１１は、たとえばＲＡＭ１４に記録されたデータから、必要な端点活性値および活性値を読み出す。制御部１１は、読み出された端点活性値および活性値について、上述のようにその通常リンクと接続された端点の端点活性値を計算する。そして制御部１１は、このように計算した端点活性値を、たとえばＲＡＭ１４に記録する。

【００８３】

ステップＳ５４においては、参照リンクに対して、そのリンクの端点の活性値を計算する処理が行われる。

【００８４】

ステップＳ５２における判別により、リンクL_{ij}は参照リンクであることが明

らかになっている。通常リンクL_{ij}に関して、エレメントE_iに接続する端点T_{ij}の端点活性値t_{ij}は、このリンクL_{ij}を除いたエレメントE_iに接続するすべてのリンクL_{ik}、L_{il}およびL_{im}の端点活性値t_{ik}、t_{il}およびt_{im}と、このエレメントE_iがリンクL_{ij}により接続するエレメントE_jの活性値e_jを加算することにより求められる。

【００８５】

エレメントE_iの端点T_{ij}の端点活性値は、端点T_{ij}を一端とするリンクL_{ij}が参照リンクの場合、リンクL_{ij}の他端が接続されているエレメントE_jの端点の端点活性値のうちそのリンクL_{ij}と接続されている端点T_{ji}を除いたすべての端点の端点活性値、およびそのリンクL_{ij}が接続されるエレメントE_jの活性値e_jの和を取ることにより得られる。

【００８６】

文書処理装置の制御部１１は、たとえばＲＡＭ１４に記録されたデータから、必要な端点活性値および活性値を読み出す。制御部１１、読み出された端点活性値および活性値について、上述のようにその参照リンクの端点活性値を計算する。そして制御部１１は、このように計算した端点活性値を、たとえばＲＡＭ１４に記録する。

【００８７】

ステップＳ５３における通常リンクの処理、およびステップＳ５４における参照リンクの処理は、ステップＳ４２のカウントのカウント値ｉにより参照されているエレメントE_iに接続するすべてのエレメントE_jについてのリンクL_{ij}に対して実行される。

【００８８】

ステップＳ５５においては、文書処理装置の制御部１１は、ステップＳ５３またはステップＳ５４での計算に基づいて、エレメントE_iの端点活性値を計算する。制御部１１は、この計算により得られた端点活性値をたとえばＲＡＭ１４に記録する。

【００８９】

ステップＳ５６においては、エレメントE_iに接続するすべてのリンクについ

て端点活性値t_{ij}が計算されたか否かが判別される。そして、すべてのリンクについて端点活性値が計算されているときには“ＹＥＳ”としてステップＳ５７に進み、すべてのリンクについて端点活性値が計算されていないときには“ＮＯ”としてステップＳ５８に進む。

【００９０】

ステップＳ５７においては、Ｓ５６にてエレメントE_iのすべてのリンクL_{ij}について端点活性値t_{ij}が求められたことが判別されたので、エレメントE_iの活性値e_iの更新を実行する。

【００９１】

エレメントE_iの活性値e_iの新たな値すなわち更新値は、エレメントE_iのすべての端点の活性値の和$e_i' = e_i + \Sigma t_j'$を取ることにより求められる。ここで、“’”は、新たな値という意味である。このように、活性値は、そのエレメントに接続するすべてのリンクについて、そのエレメントに接続する端点の端点活性値の総和となる。

【００９２】

文書処理装置の制御部１１は、たとえばＲＡＭ１４に記録されたデータから必要な端点活性値t_{ij}を読み出す。制御部１１は、上述したような計算を実行し、そのエレメントE_iの活性値e_iを算出する。そして、制御部１１は、計算した新たな活性値e_iをたとえばＲＡＭ１４に記録する。

【００９３】

次に、上述した活性値に基づいて行う語義の関連度の計算について、図１１に示すフローチャートを参照して説明する。語義の関連度の計算は、図４および図６に示す処理を行う前にあらかじめ行う前処理であるから、一度実行すればよい。

【００９４】

最初のステップＳ６１において、制御部１１は、電子辞書内の語の語義の説明を用い、辞書を使って語義のネットワークを作成する。すなわち、辞書における各語義の説明と、この説明中に現れる語義との参照関係から、上述したような語義のタグ付けによる構造のネットワークを作成する。これは、最上位のエレメン

トを辞書として、図２に示したようなタグ付けによる内部構造を構成することに相当する。制御部１１は、ＲＡＭ１４に記録した語義とその説明を順に読み出して、ネットワークを作成する。制御部１４は、このようにして作成した語義のネットワークをたとえばＲＡＭ１４や記録／再生部３１に記録する。

【００９５】

なお、この辞書は、たとえば通信回線から受信部２１にて受信することができる。また、たとえばＣＤ－ＲＯＭなどの記録媒体３２によって提供され、記録／再生部３１で再生することができる。

【００９６】

ステップＳ６２において、ステップＳ６１で作成された語義のネットワーク上で、上述した活性拡散を行う。この活性拡散により、各語義の活性値は、上記辞書により与えられたタグ付けによる内部構造に応じて更新される。

【００９７】

ステップＳ６３においては、ステップＳ６１で作成された語義のネットワークを構成する一つの語義s_iを選択し、ステップＳ６４においては、この語義s_iに対応する語彙エレメントE_iの活性値e_iの初期値を適当の変化させ、このときの活性値の差分Δe_iを計算する。

【００９８】

ステップＳ６５においては、ステップＳ６４におけるエレメントE_iの活性値e_iの初期値の変化に対応する、語義s_jに対応するエレメントE_jの活性値e_jの差分Δe_jを求める。ステップＳ６６においては、ステップＳ６５で求めた差分Δe_jをステップＳ６４で求めたΔe_iで除した商$\Delta e_j / \Delta e_i$を、語義s_iの語義s_jに対する関連度とする。ある語義の活性値をステップＳ６４で変えたのに応じて、関連する語の活性値の変わることとなる。

【００９９】

ステップＳ６７においては、語義s_iとs_jとのすべての組について関連度の演算が終了したか否かについて判断する。そして、すべての語義の組について関連度の演算が終了したときには“ＹＥＳ”として、この一連の処理を終了する。すべての語義の組について関連度の演算が終了していないときには、“ＮＯ”とし

て、ステップＳ６３にもどり、関連度の演算が終了していない組について関連度の演算を継続する。

【０１００】

このように計算された関連度は、図１２に示すように、それぞれの語義と語義の間に定義される。この語義の表においては、関連度は正規化され、０から１までの値をとる。すなわち、この語義の表においては“コンピュータ”、“テレビ”、“ＶＴＲ”の間の相互の関連度が示されている。“コンピュータ”と“テレビ”の関連度は０．５５、“コンピュータ”と“ＶＴＲ”の関連度は０．２５、“テレビ”と“ＶＴＲ”の関連度は０．６０である。制御部１１は、このように作成した関連度をたとえばＲＡＭ１４に記憶する。

【０１０１】

ステップＳ６３からステップＳ６７のループにおいては、制御部は、必要な値をたとえばＲＡＭ１４や記録／再生部３１から順に読み出して、上述したように関連度を計算する。制御部１１は、計算した関連度をたとえばＲＡＭ１４や記録／再生部３１に記録する。

【０１０２】

次に、上述したように算出された関連度を用いた文書分類について説明する。この関連度を利用した文書分類は、先に説明した図５のＧＵＩにおける文書分類に用いられる。

【０１０３】

関連度による文書分類は、各分類項目の特徴を示す分類モデルを参照して関連度に基づいて行われる。分類モデルとは、各分類項目に特徴的な、固有名詞、固有名詞以外の語義、アドレスなどを含んで構成される。たとえば、図１３に示す分類モデルは、各分類項目、いわゆるカテゴリに対して、固有名詞、固有名詞以外の語義、アドレスの欄を有している。この分類モデルにおいては、分類項目は“スポーツ”、“社会”、“コンピュータ”、“植物”、“美術”および“美術”の各項目から構成されている。これらの分類項目に対応する固有名詞として、“Ａ氏”、“Ｂ社”、“Ｃ社”および“Ｇ社”、“Ｄ種”、“Ｅ氏”、“Ｆ氏”がそれぞれ示されている。上記分類項目に対応する固有名詞以外の語義として、

“野球”および“グランド”、“労働”および“雇用”、“モバイル”、“桜１”および“オレンジ１”、“桜２”および“オレンジ２”、“桜３”がそれぞれ示されている。また、上記分類項目に対応するアドレスとして、“１２３４５”、“２２２２２”、“３３３３３”、“４４４４４”、“５５５５５”、“６６６６６”がそれぞれ示されている。なお、“桜１”、“桜２”および“桜３”は“桜”の第１の語義(１１１１１)、第２の語義(１１１１２)および第３の語義(１１１１３)を示している。また、“オレンジ１”および“オレンジ２”は、“オレンジ”の第１および第２の語義を示している。

【０１０４】

各分類項目の分類モデルは、タグ付けによる内部構造による活性値に基づいて抽出される。上述したように、文書処理装置の制御部１１は、ステップＳ３２において活性値が所定の閾値を超えるエレメントを抽出し、ステップＳ３３においてこのエレメントからすべての固有名詞を取り出してインデックスに加え、ステップＳ３４においては固有名詞以外の語義を取り出してインデックスに加える。このように分類モデルの特徴の欄は、たとえば上述の手順により生成されたインデックスを分類項目ごとにまとめたものである。

【０１０５】

図６におけるステップＳ２３で行われる文書の自動分類は、このような分類モデルを参照して、図１４のフローチャートに示す一連の手順に従って、語義の関連度に基づいて行われる。

【０１０６】

ステップＳ７１においては、制御部１１は、分類モデルの各分類項目C_iに含まれる固有名詞の集合と、ステップＳ６２において文書から抽出されインデックスに入れられた語のうちの固有名詞の集合とについて、これらの共通集合の数を$P(C_i)$とする。そして、制御部１１は、このようにして算出した数$P(C_i)$をたとえばＲＡＭ１４に記録する。

【０１０７】

ステップＳ７２においては、制御部１１は、その文書のインデックス中の語義と各分類項目Ｃｉに含まれる語義との関連度を図１２の語義の表を参照し、語義

の関連度の総和Ｒ（C_i）を演算する。すなわち、制御部１１は、分類モデルにおける固有名詞以外の語について、ステップＳ６１で算出した関連度の総和Ｒ（C_i）をとる。そして、制御部１１は、算出した関連度の総和Ｒ（C_i）をたとえばＲＡＭ１４に記録する。

【０１０８】

ステップＳ７３においては、項目C_iに対する文書の関連度を、

$$Rel(C_i) = mP(C_i) + nR(C_i)$$

と定義する。ここで、係数ｍ、ｎは定数で、それぞれの値の関連度への貢献の度合いを表すパラメータである。制御部１１は、ステップＳ３３で算出した共通集合の数Ｐ（C_i）およびステップＳ６４で算出した語義の関連度の総和Ｒ（C_i）をたとえばＲＡＭ１４から読み出し、上述の式に当てはめて文書の関連度Ｒｅｌ（C_i）を算出する。なお、これらの係数ｍ、ｎの値としては、たとえばｍ＝１０、ｎ＝１とすることができる。そして、制御部１１は、このように算出した文書の関連度Ｒｅｌ（C_i）をたとえばＲＡＭ１４に記録する。

【０１０９】

係数ｍおよびｎの値は、統計的手法を使って推定することもできる。すなわち、制御部１１は、複数の係数ｍおよびｎの組について文書の関連度Ｒｅｌ（C_i）が与えられると、上記係数を最適化により求めることができる。

【０１１０】

ステップＳ７４においては、制御部１１は、項目C_iに対する関連度Ｒｅｌ（C_i）が全項目中最大で、その関連度の値がある閾値を超えているとき、分類項目Ｃｉに文書を分類する。すなわち、制御部は、複数の項目についてそれぞれ文書の関連度Ｒｅｌ（C_i）を作成し、最大の関連度Ｒｅｌ（C_i）が閾値を超えているときには、文書を上記項目に分類C_iする。最大の関連度Ｒｅｌ（C_i）が閾値を超えていないときには、文書の分類は行わない。

【０１１１】

このように、文書中に含まれる語義間の関連度の計算とそれに基づく文書の分類の手順は、複数のエレメントから構成されるタグ付けによる内部構造を有する文書を処理し、この文書を複数の分類項目の内の一つの分類項目に分類する。す

なわち、この手順は、文書と各分類項目との関連度を算出し、算出された関連度に基づいて上記文書を分類する分類項目を決定する。

【０１１２】

ここで、文書を分類する分類項目は、文書から抽出された固有名詞および／または語義を含む分類モデルによって特徴づけられる。そして、このような分類モデルを用いて、各分類項目の分類モデルに含まれる固有名詞および文書から抽出された固有名詞についての共通の数を算出し、各分類項目の上記分類モデルに含まれる語義に対する上記文書の関連度の総和を算出する。さらに、文書に含まれる固有名詞と関連度に基づいて抽出された固有名詞おいて重複する固有名詞の数と、語義の関連度の総和とに基づいて上記文書を分類する分類項目を決定する。この語義の関連度は、上述したようなタグ付けによる内部構造に基づいて決定される。

【０１１３】

文書の分類は、たとえば共通する固有名詞の数および語義の関連度の線形結合が最大となって、所定の閾値を越えるような項目に対して行われる。このような共通する固有名詞の数および語義の関連度の線形結合の係数は、文書と分類項目の関連の大きさから、上述したように統計的に決定することができる。

【０１１４】

次に、文書処理装置の記録／再生部３１において記録／再生される記録媒体３２について説明する。記録媒体には、複数のエレメントからタグ付けによる内部構造を有する文書を処理する文書処理プログラムが記録されている。この記録媒体３２としては、情報の記録／再生が可能なたとえばフロッピーディスクが利用される。

【０１１５】

記録媒体３２において、文書処理プログラムは、エレメントの最小単位である語義ごとに他の語義を参照する辞書を用い、語義の参照関係を組織する参照関係組織処理と、参照関係組織処理で組織された参照関係の組織の構造に基づいて語義にそれぞれ活性値を付与する活性値付与処理と、活性値付与処理で上記語義に付与された活性値に、参照関係の組織の構造に基づいた演算を施すことにより、

語義に新たに活性値を付与する活性値付与処理と、活性値付与処理で一の語義に付与された活性値を独立変数とし、活性値演算処理で他の語義に付与された活性値を従属変数とし、活性演算処理で他の語義に付与された活性値の微分を活性値付与工程で一の語義に付与された活性値の微分で除した微分商を一の語義と他の語義の関連度として演算する関連度演算処理との各処理工程を有するものである。

【０１１６】

記録媒体３２においては、参照関係組織処理で用いられる辞書の各語義にはその語義の属性を示す属性情報が付与され、上記参照関係の組織は上記属性情報に基づいて作成される。

【０１１７】

記録媒体３２において、文書処理プログラムは、エレメントの最小単位である語義の間の相互の関連度を算出する関連度算出処理と、文書を分類する複数の分類項目について、分類項目の特徴をあらわす語義を含んでなる分類モデルを用い、各分類モデルの含む語義との関連度に基づいて文書を分類する文書分類処理との各処理工程を有する。

【０１１８】

記録媒体３２において、関連度算出処理は、エレメントの最小単位である語義ごとに他の語義を参照する辞書を用い、上記語義の参照関係を組織する参照関係組織処理と、参照関係組織処理で組織された参照関係の組織の構造に基づいて語義の構造に基づいた演算を施すことにより、語義に新たに活性値を付与する活性値付与処理と、活性値付与処理で一の語義に付与された活性値を独立変数とし、活性値演算処理で他の語義に付与された活性値を従属変数とし、活性値演算処理で他の語義に付与された活性値の微分を活性値付与処理で一の語義に付与された活性値の微分で除した微分商を一の語義と他の語義の関連度として演算する関連度演算処理との各処理工程を有する。

【０１１９】

なお、本実施の形態においては、文書へのタグ付けの方法の一例を示したが、本発明がこのタグ付けの方法に限定されないことはもちろんである。また、本実

施の形態においては、文書処理装置の受信部２１に外部から文書が送信されるとしたが、本発明はこれに限定されない。たとえば、上記文書は、文書処理装置のＲＯＭ１３に書き込まれていたり、記録／再生部３１において記録媒体３２から読み出されてもよい。

【０１２０】

また、上述の実施の形態においては、文書処理装置の表示部３０に表示された文書から所望のエレメントを選択するデバイスとしてマウスを例示したが、本発明がこれに限定されないことはいうまでもない。文書処理装置におけるエレメントの入力には、タブレット、ライトペン等の他のデバイスを利用することができる。

【０１２１】

【発明の効果】

上述のように、本発明に係る文書処理方法および装置ならびに記録媒体を利用することにより、語義の関連度を算出することができる。この語義の関連度を利用することにより、語義の関連度に基づいて、ユーザの興味を反映した文書の自動分類のような文書処理を実行することができるようになる。語義の関連度に基づく文書処理は、自動的に実行することができるので、文書処理の際のユーザの負担を軽減する。

【図面の簡単な説明】

【図１】

本実施の形態を適用した文書処理装置の構成を示すブロック図である。

【図２】

文書のタグ付けによる内部構成を示す図である。

【図３】

文書のタグ付けによる内部構成を表示したウィンドウを示す図である。

【図４】

本実施の形態を適用した文書処理装置の動作を示すフローチャートである。

【図５】

文書の自動分類を行うＧＵＩを示す図である。

【図６】

文書を自動分類するフローチャートである。

【図７】

文書の特徴を発見してインデックスを作成するフローチャートである。

【図８】

活性拡散を示すフローチャートである。

【図９】

活性拡散の処理を説明する図である。

【図１０】

活性拡散のリンク処理のフローチャートである。

【図１１】

語義の関連度の計算のフローチャートである。

【図１２】

語義の関連度の表を示す図である。

【図１３】

分類モデルの表を示す図である。

【図１４】

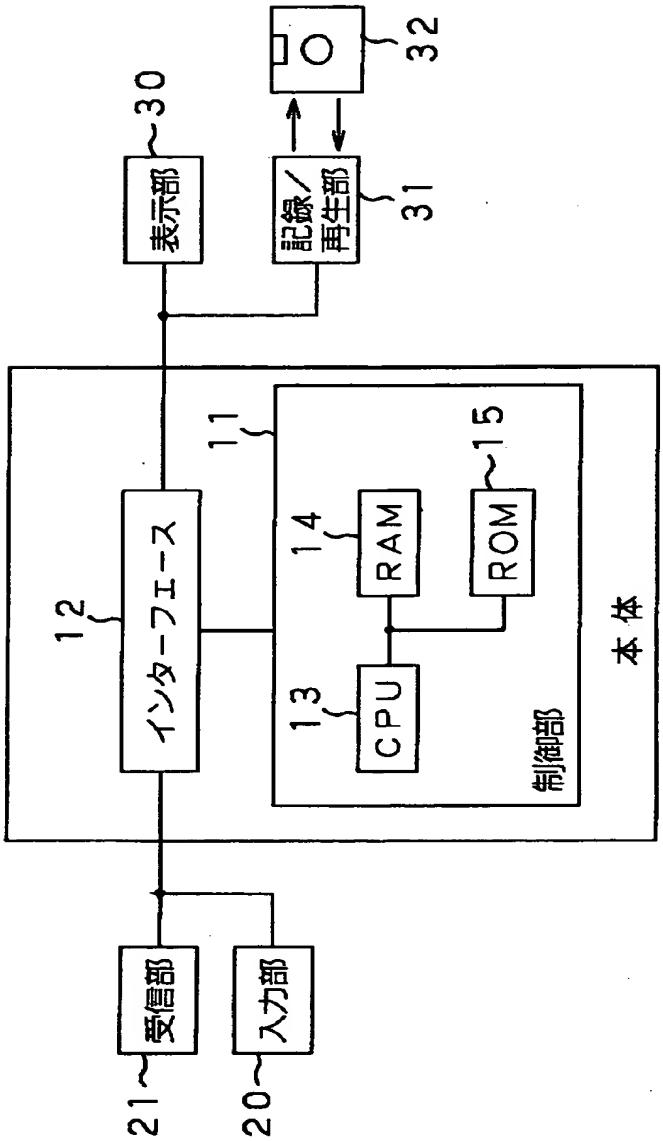
関連度による文書分類のフローチャートである。

【符号の説明】

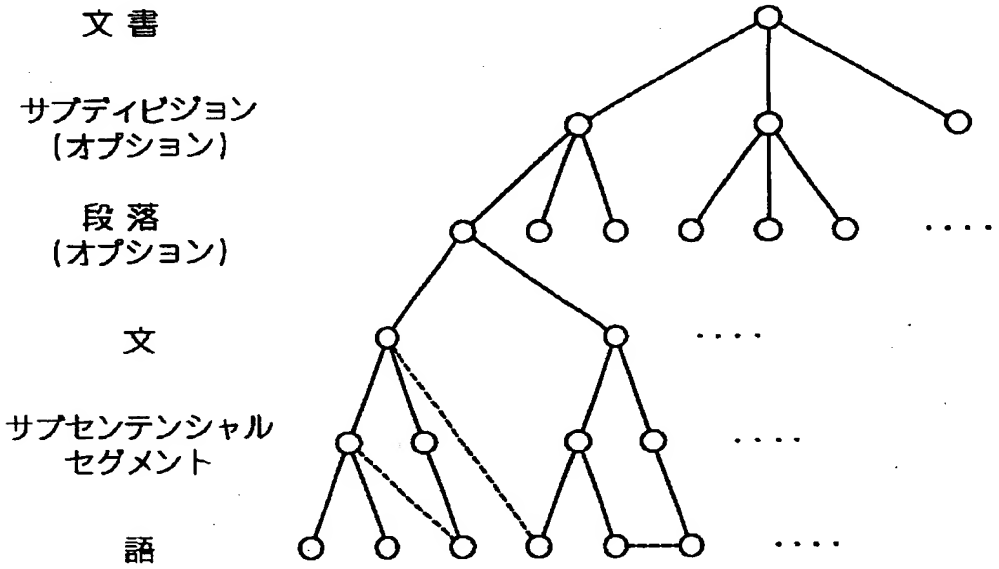
１０　本体、１１　制御部、１２　インターフェース、１３　ＣＰＵ、２０　入力部、２１　受信部、３０　表示部、３１　記録／再生部

【書類名】　図面

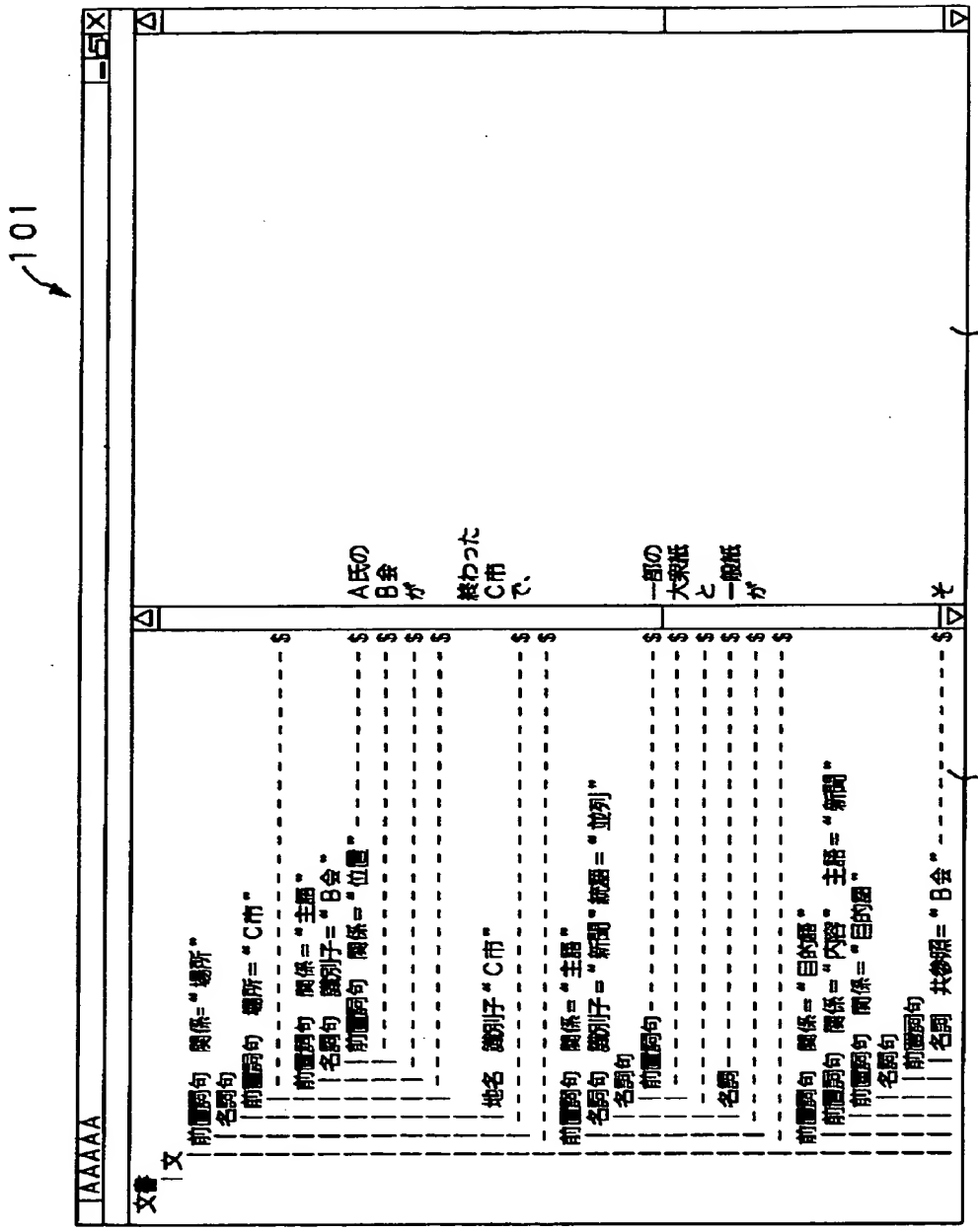
【図１】


32
30
表示部
記録／再生部
31
11
15
ROM
14
RAM
12
インターフェース
13
CPU
制御部
本体
受信部
入力部
21
20

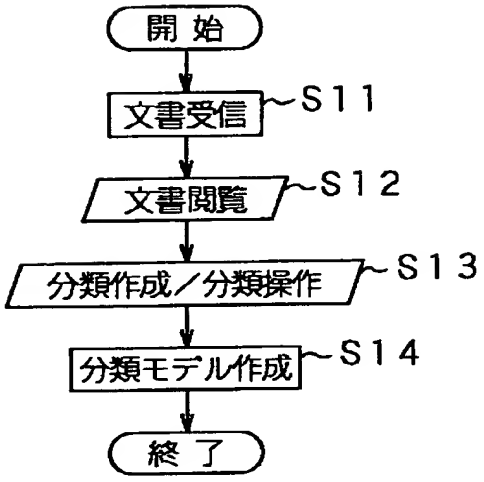
【図２】



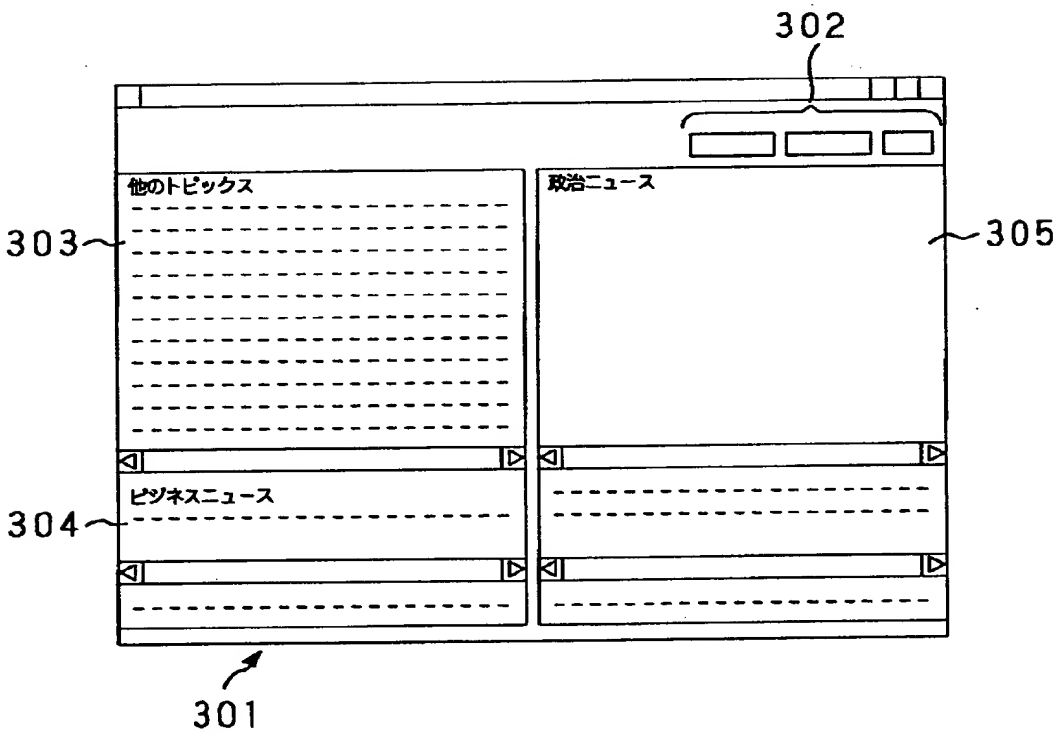
【図３】


101
103
102
AAAAA
文書
文
前置詞句 関係="場所"
名詞句
前置詞句 場所="C市"
前置詞句 関係="主語"
名詞句 識別子="B会"
前置詞句 関係="位置"
地名 識別子"C市"
前置詞句 関係="主語"
名詞句 識別子="新聞" 修飾="並列"
名詞句
前置詞句
名詞
前置詞句 関係="目的語"
前置詞句 関係="内容" 主語="新聞"
前置詞句 関係="目的語"
名詞句
前置詞句
名詞 共参照="B会"
A氏の
B会
が
終わった
C市
で、
一部の
大衆紙
と
一般紙
が
。

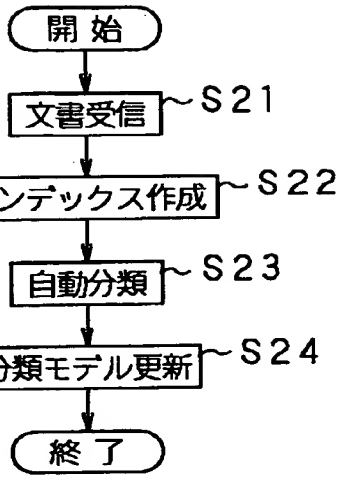
【図４】


開 始
文書受信
S11
文書閲覧
S12
分類作成／分類操作
S13
分類モデル作成
S14
終 了

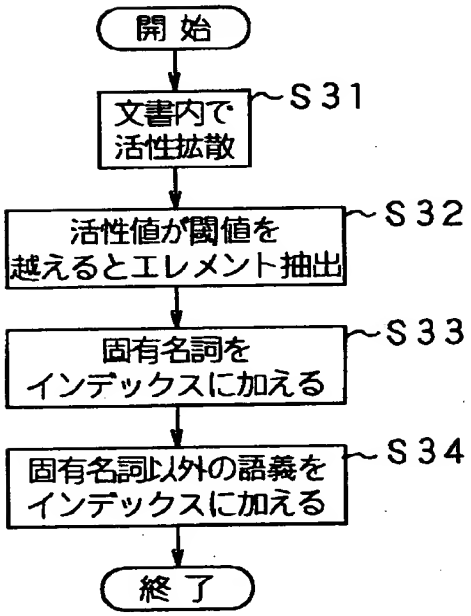
【図５】


302
他のトピックス
303
政治ニュース
305
ビジネスニュース
304
301

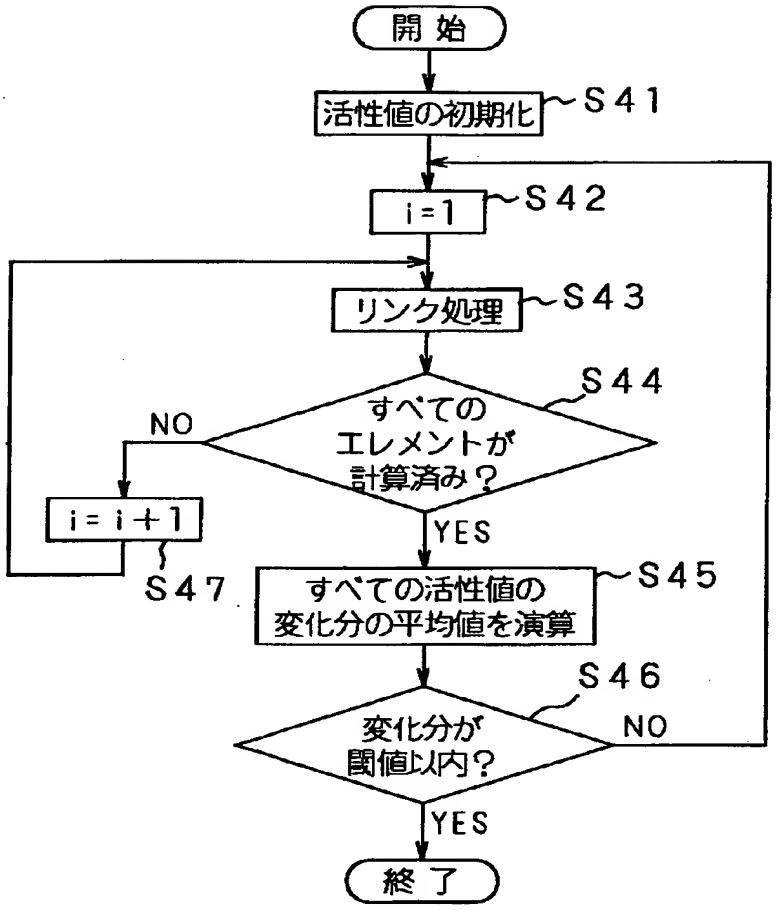
【図６】


開 始
文書受信
S21
インデックス作成
S22
自動分類
S23
分類モデル更新
S24
終 了

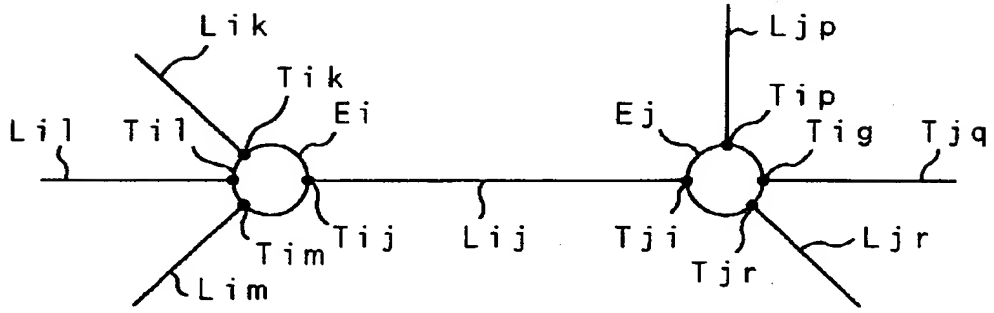
【図７】


開 始
文書内で
活性拡散
S31
活性値が閾値を
越えるとエレメント抽出
S32
固有名詞を
インデックスに加える
S33
固有名詞以外の語義を
インデックスに加える
S34
終 了

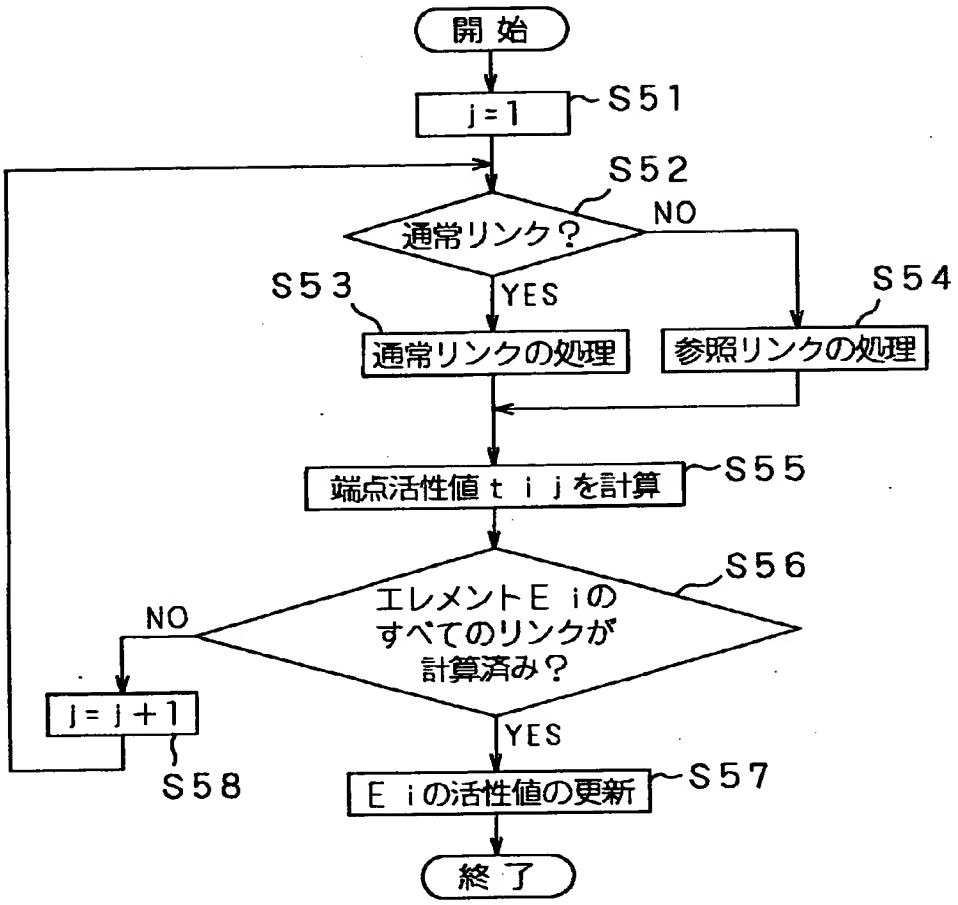
【図８】


開 始
活性値の初期化
S41
i=1
S42
リンク処理
S43
S44
すべての
エレメントが
計算済み？
NO
YES
i=i+1
S47
すべての活性値の
変化分の平均値を演算
S45
S46
変化分が
閾値以内？
NO
YES
終 了

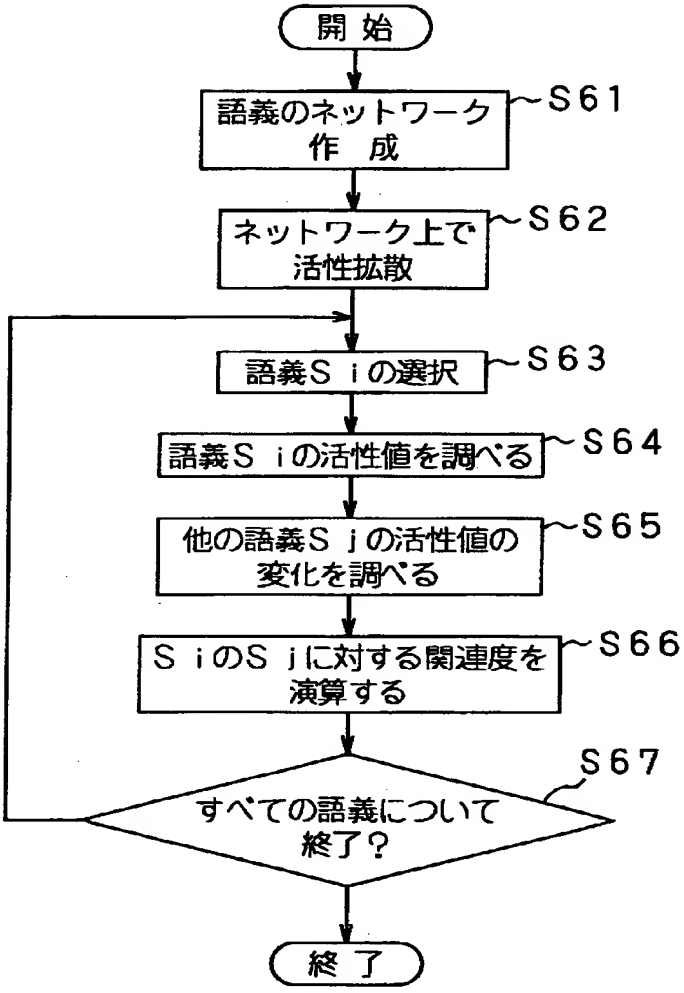
【図９】


Lik
Tik
Ei
Lil
Til
Tim
Tij
Lij
Lim
Ljp
Tip
Ej
Tig
Tjq
Tji
Tjr
Ljr

【図１０】


開 始
j=1
S51
S52
通常リンク？
NO
YES
S53
S54
通常リンクの処理
参照リンクの処理
端点活性値ｔｉｊを計算
S55
S56
エレメントＥｉの
すべてのリンクが
計算済み？
NO
YES
j=j+1
S58
Ｅｉの活性値の更新
S57
終 了

【図１１】



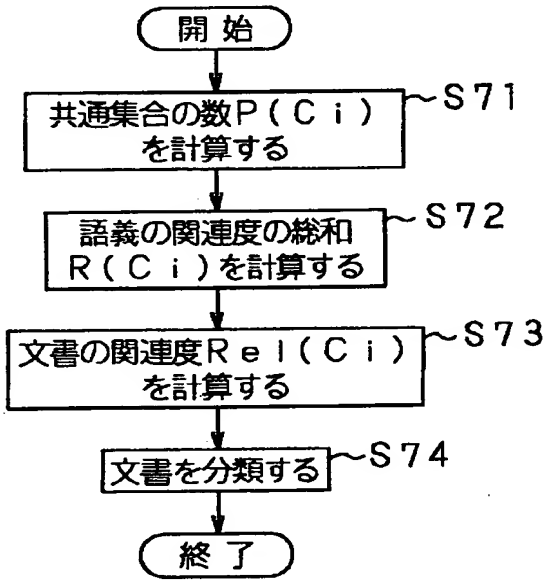
【図１２】

	コンピュータ	テレビ	
コンピュータ		0.55	
テレビ	0.55		
VTR	0.25	0.60	

【図１３】

分類項目	スポーツ	社 会	コンピュータ	植 物	美 術	イベント
固有名詞	A 氏	B 社	C社 G社	D 種	E 氏	F 氏
語 義	野 球 グランド	労働 雇用	モバイル	桜1 (11111) オレンジ1	桜2 (11112) オレンジ2	桜3 (11113)
文書 アドレス	12345	22222	33333	44444	55555	66666

【図１４】



【書類名】　要約書

【要約】

【課題】　語義の関連度に基づいて文書を分類する。

【解決手段】　複数のエレメントから構成され、タグ付けによる内部構造を有する文書を処理する文書処理装置であって、受信部２１で受信した複数の文書をたとえば本体の制御部１１のＲＡＭ１４に記憶し、ＣＰＵ１３はたとえばＲＯＭに記録された手順にしたがって、文書の特徴を表す特徴情報を抽出し、文書を分類する分類モデルを構成する複数の分類項目について、抽出した文書の特徴情報と上記分類項目ごとの特徴情報との関連度に応じて、各文書を分類項目に分類し、その結果を表示部３０に表示する。

【選択図】　図１

認定・付加情報

特許出願の番号	平成１１年　特許願　第０１３３０７号
受付番号	５９９０００４９７７４
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１１年　１月２５日

＜認定情報・付加情報＞

【提出日】　平成11年 1月21日

次頁無

出　願　人　履　歴　情　報

識別番号　　〔０００００２１８５〕

１．変更年月日　　１９９０年　８月３０日
［変更理由］　新規登録
住　所　東京都品川区北品川６丁目７番３５号
氏　名　ソニー株式会社